International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

AI & NLP-Based Question Generation Using the T5 Model for Education

Balamurugan A[1], Bharathidasan S[2], Arputha Joshwa B[3], Rathinapriya V[4], Balaji V[5]

[1,2,3,4,5]Department of Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai, Tamilnadu, India.

Email ID: *balamuruganm4eo@gmail.com[1], bharathidasansm@gmail.com[2], arputhajoshwa@gmail.com[3], rathinapriyavasu02@gmail.com[4], balajivishnu040@gmail.com[5]*

Abstract

With the increasing demand for efficient and scalable educational assessments, Artificial Intelligence (AI) and Natural Language Processing (NLP) have emerged as transformative tools. Automated Question Generation (AQG) is a key application that leverages AI to generate relevant and diverse questions from textual content. This study explores the implementation of AQG using the Text-To-Text Transfer Transformer (T5) model, a state-of-the-art deep learning model for NLP tasks. The proposed system is fine-tuned on educational datasets to enhance question fluency, relevance, and accuracy. Unlike traditional rule-based approaches, the T5 model can dynamically generate factual, inferential, and multiple-choice questions across various subjects. This research aims to improve assessment quality, reduce the manual workload of educators, and support adaptive learning environments. The findings highlight the model's potential in creating high-quality questions while addressing challenges such as contextual accuracy, grammatical correctness, and bias mitigation. Future work will focus on refining question quality, integrating AQG into intelligent tutoring systems, and enhancing contextual understanding.

Keywords: *Automated Question Generation, AI in Education, Natural Language Processing, T5 Model, Deep Learning, Question Answering, Educational Technology, Adaptive Learning Systems, Machine Learning, Text Generation, Contextual Understanding.*

1. Introduction

In the digital era, the integration of Artificial Intelligence (AI) and Natural Language Processing (NLP) in education has revolutionized traditional learning methodologies. [1] One of the most promising applications of AI in education is Automated Question Generation (AQG), which plays a crucial role in enhancing assessments, personalized learning, and content development. Traditionally, educators manually create questions for tests, quizzes, and assignments—a process that is often time-consuming and prone to subjectivity. With the advent of AI-driven AQG, the question-generation process has become more efficient, scalable, and adaptable, catering to a wide range of learners and educational contexts.[1] This paper explores an AI-based approach to question generation using the T5 (Text-To-Text Transfer Transformer) model, a powerful deep learning framework designed for various text-based tasks. [1]Unlike rule-based question generation techniques that rely on predefined templates, T5 leverages deep learning to generate diverse, contextually relevant, and grammatically sound questions from textual content. This advancement has the potential to reduce the manual workload on educators, improve the quality of assessments, and create adaptive learning environments where students receive tailored questions based on their understanding and progress. [1] Assessment is a fundamental aspect of education, helping educators gauge students' knowledge, comprehension, and critical thinking abilities. However, the traditional process of crafting assessments manually is not only

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

labor-intensive but also limited in scalabilit. In large classrooms, online courses, and self-learning platforms, there is a pressing need for automated solutions that can generate questions dynamically and efficiently. [1]

1.1. Automated Question Generation Addresses Several Key Challenges in Education

Scalability: AI can generate a vast number of questions across multiple subjects and difficulty levels without additional manual effort.

Personalization: AI-driven AQG enables personalized learning by adapting questions to individual student needs and knowledge levels.

Efficiency: Automating the question-generation process allows educators to focus on teaching and student engagement rather than repetitive assessment creation.

Content Diversity: AI can generate multiple types of questions (e.g., multiple-choice, fill-in-the-blanks, short-answer, and inferential questions) to assess different cognitive skills.

Reduction of Bias: Unlike human-generated assessments, which may introduce unintended bias, AI models trained on diverse datasets can produce more balanced and objective questions.

1.2. T5 Model for Question Generation

The T5 (Text-To-Text Transfer Transformer) model, developed by Google, is a transformer-based NLP model designed for a wide range of text-based tasks, including machine translation, text summarization, and question generation. Unlike conventional question-generation models, T5 treats every NLP problem as a text-to-text task, making it highly flexible and adaptable. By fine-tuning the T5 model on large datasets containing question-answer pairs, it learns to generate high-quality questions that align with the context, grammar, and semantics of the given input text. This model surpasses traditional rule-based and retrieval-based methods by offering better contextual understanding, greater diversity in question types, and adaptability to various domains.

2. Literature Survey

Automated Question Generation (AQG) is an emerging area in Artificial Intelligence (AI) and Natural Language Processing (NLP) that aims to automate the creation of assessment questions. Traditional question generation methods rely on manual efforts or rule-based approaches, which can be time-consuming and inconsistent. Recent advancements in deep learning models like T5, BERT, and GPT have significantly improved the ability to generate contextually relevant and diverse questions.

2.1. Objective of Automated Question Generation (AQG)

The objective of Automated Question Generation (AQG) is to revolutionize the way questions are formulated in educational and assessment contexts. [2]The primary aim is to automate the process of generating questions that are contextually accurate, syntactically sound, and pedagogically meaningful. One of the key goals of AQG is to enhance learning efficiency by reducing the manual effort required from educators in preparing assessments. [2] By automating this process, AQG enables the rapid generation of diverse and well-structured questions, ensuring a more efficient workflow in academic and professional settings. [2]Another crucial objective is to improve personalized learning. AI-driven AQG allows for the dynamic adaptation of questions to individual students' learning levels, making education more engaging and effective. By generating questions that align with the learner's progress, AQG can support adaptive learning models that modify instructional content based on student performance. Furthermore, AQG aims to support multiple question formats, ensuring that assessments are varied and comprehensive. Unlike traditional rule-based methods, which often rely on rigid templates, AI-based AQG models can generate factual, inferential, open-ended, and multiple-choice questions from textual content. This enables educators to assess students across different cognitive levels, from knowledge recall to analytical thinking. [2] An additional objective is to reduce bias in question generation. Since AI models can be trained on diverse and inclusive datasets, they can help minimize unintended biases that often arise in manually created assessments. This contributes to the creation of fair and balanced question banks that cater to learners from varied backgrounds. Lastly, AQG

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

seeks to enhance the capabilities of intelligent tutoring systems (ITS). By integrating AQG into digital learning platforms, educators can create automated assessments that respond dynamically to student needs. This facilitates real-time feedback and knowledge assessment, fostering a more interactive learning experience.

2.2. Scope of AQG Research and Applications

The scope of AQG research and applications spans across various domains, including education, corporate training, intelligent tutoring systems, AI-powered chatbots, and medical education. The growing need for automated assessments, personalized learning experiences, and scalable evaluation methods has made AQG a critical area of study in AI and NLP.

2.2.1. Education and E-Learning

The most prominent application of AQG is in academic settings, where AI-based systems help generate customized quizzes, assignments, and self-assessments. Online learning platforms, such as Coursera, Udemy, and edX, leverage AI to automate question generation based on course material. This enables a scalable and efficient approach to knowledge assessment, allowing educators to focus more on curriculum development rather than manual test creation.

2.2.2. Intelligent Tutoring Systems (ITS)

AI-powered Intelligent Tutoring Systems (ITS) use AQG to create real-time assessments and feedback loops. These systems dynamically adjust question difficulty based on student performance, ensuring an adaptive learning experience. Research in ITS has shown that automated assessments improve student retention and engagement by providing interactive and personalized content.

2.2.3. Corporate Training and Skill Development

In the corporate sector, AQG is widely used in training programs, professional certification exams, and skill assessment modules. AI-generated questions allow organizations to conduct standardized evaluations without the need for constant human intervention. By using AQG, companies can streamline their employee training programs, ensuring consistency in assessment quality and effectiveness.

2.2.4. AI-Powered Chatbots and Virtual Assistants

Another significant application of AQG is in AI-powered chatbots and virtual assistants. Many customer service platforms incorporate automated FAQ generation using NLP models to improve user experience and query resolution. In education, virtual tutors use AQG to provide students with instant practice questions, making learning more interactive and engaging.

2.2.5. Medical Education and Healthcare

AQG is also transforming medical education by assisting in the creation of clinical case-based questions for medical students. AI-generated questions help simulate real-world scenarios, making it easier for students to develop critical thinking and diagnostic skills. Additionally, AQG is being explored in the development of self-assessment tools for medical licensing exams.

2.2.6. Computational Linguistics and NLP Research

Beyond direct applications in education, AQG contributes to broader NLP research in areas such as text summarization, information retrieval, and semantic understanding. The advancements in transformer models like T5, GPT, and BERT have improved context-aware question generation, leading to better language models for dialogue systems and AI tutors.

3. Proposed Methodology

To develop an effective Automated Question Generation (AQG) system, we propose a deep learning-based approach utilizing the T5 (Text-To-Text Transfer Transformer) model. [3] Unlike traditional rule-based methods, our approach leverages contextual learning and sequence-to-sequence processing to generate grammatically accurate and contextually relevant questions from input text. The system is designed to handle various question formats, including factual, inferential, and multiple-choice questions, making it adaptable for different educational contexts. The methodology consists of several stages: data collection, preprocessing, model fine-tuning, question generation, and evaluation. Initially, a large dataset of

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

text passages and corresponding question-answer pairs is curated and cleaned to ensure quality input. The T5 model is then fine-tuned on this dataset to enhance its ability to generate well-structured questions. During inference, the model processes textual input, extracts key concepts, and formulates relevant and diverse questions based on learned patterns. The generated questions undergo quantitative evaluation using precision, recall, and F1-score, along with qualitative human assessments to validate fluency and relevance. [3]

3.1. Architecture Diagram

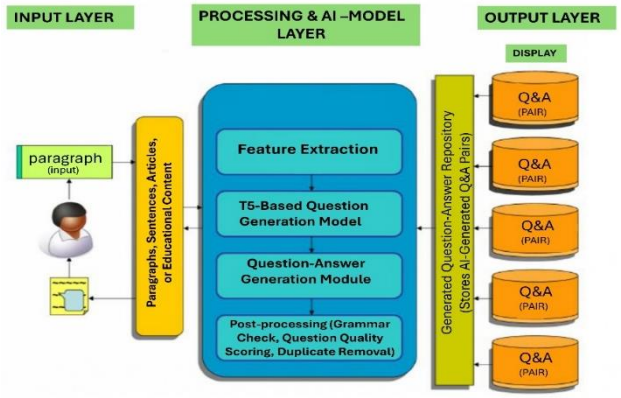

Figure 1 System Architecture Diagram

Efficient Architecture for AI-Based Automated Question Generation (AQG) Using T5: The proposed architecture for the Automated Question Generation (AQG) system is designed to be efficient, scalable, and adaptable for various educational applications. The system is built around the T5 (Text-To-Text Transfer Transformer) model, which processes textual input and generates contextually relevant questions. The architecture consists of multiple interconnected components, each responsible for a critical stage in the question-generation pipeline. [3] The first stage is the Data Collection and Preprocessing Module, where raw textual data is gathered from diverse sources, including educational repositories, textbooks, and online learning platforms. The text is cleaned, normalized, tokenized, and structured to enhance readability and ensure consistency. This module also incorporates Named Entity Recognition (NER) and semantic parsing techniques to extract key concepts from the text, which helps in generating meaningful questions. [3] Next, the Feature Extraction & Encoding Module converts preprocessed text into tokenized embeddings using the SentencePiece tokenizer (specific to T5). This stage ensures that the text is transformed into a format that the deep learning model can effectively process. Positional encoding is applied to retain sentence structure, and optional knowledge graph mapping is integrated to improve domain-specific question generation. The core of the architecture is the T5-Based Question Generation Module, which is responsible for generating diverse and grammatically accurate questions. The model is fine-tuned on question-answer pairs from educational datasets, allowing it to learn different types of question formats, including factual (Who, What, When), inferential (Why, How), and multiple-choice questions. The system utilizes attention mechanisms to focus on important parts of the text while generating questions and employs beam search and top-k sampling to ensure fluency and variation in outputs. [4]To enhance reliability, the Evaluation & Filtering Module validates the generated questions using both AI-based quality checks and human expert assessments. The Grammar Checker (GPT-based) ensures syntactic correctness, while the Semantic Similarity Checker (BERT-based) eliminates duplicate or irrelevant questions. A diversity score calculator measures the variety in question types, ensuring that the system does not generate repetitive content. Additionally, a human-in-the-loop validation process is incorporated, where expert reviewers periodically assess question quality and provide feedback for model improvements. Finally, the Deployment & API Integration Module enables real-world implementation of the AQG system. The trained model is deployed as a REST API using Flask or FastAPI, allowing seamless integration with learning management systems (LMS), intelligent tutoring systems, and educational apps. A PostgreSQL or NoSQL database stores generated questions for later retrieval, and a web-based UI or mobile application allows users to interact with the system in an intuitive manner. The architecture supports cloud deployment (AWS, GCP, Azure) with containerized microservices

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

(Docker/Kubernetes) to ensure scalability, high availability, and efficient workload management. [4] [4]To optimize efficiency, the system employs model pruning and quantization techniques to reduce computational overhead while maintaining high accuracy. ONNX runtime and TensorRT acceleration are used for faster inference, ensuring real-time question generation. Security measures, including data sanitization pipelines and fairness-aware training techniques, help mitigate biases in generated content. Additionally, multilingual support is planned for future improvements, enabling the system to generate questions in multiple languages for global accessibility. [4] Overall, this architecture is designed to provide fast, reliable, and high-quality question generation for automated assessments, online learning platforms, and AI-driven tutoring systems. By integrating deep learning, NLP, and cloud-based optimizations, this system enhances the efficiency of educational content creation and assessment automation, making AI-powered question generation a valuable tool in modern learning environments.

3.2. Explanation of the Proposed Model

The proposed methodology for Automated Question Generation using AI and NLP is structured into multiple phases, each playing a crucial role in enhancing the efficiency and accuracy of generated questions. Our model is based on the T5 (Text-To-Text Transfer Transformer), a state-of-the-art deep learning framework designed to handle various NLP tasks using a text-to-text format. Below is a detailed breakdown of the methodology

3.2.1. Data Collection and Preprocessing

[9] The first step in the AQG process is gathering a high-quality dataset consisting of text passages and corresponding question-answer pairs from diverse educational sources. The collected data is then preprocessed to ensure:

- Removal of noise such as irrelevant characters and duplicate entries.
- Tokenization of text to convert sentences into word embeddings.
- Normalization of text to standardize variations in spelling and formatting.
- Filtering and annotation to classify data into different categories of question types. [9]

3.2.2. Fine-Tuning the T5 Model

Once the dataset is preprocessed we fine-tune the T5 model using supervised learning. The model is trained on input text passages, where it learns to generate meaningful questions. The fine-tuning process involves, Pretraining on large-scale question-answering datasets to improve contextual understanding. Custom fine-tuning using domain-specific educational datasets to optimize performance for different subjects. Data augmentation techniques to diversify the training samples and prevent overfitting.

3.2.3. Question Generation Process

During inference, the model takes input text and generates questions like Identifying key entities, concepts, and relationships in the text Determining the most suitable question format Generating a grammatically accurate and contextually relevant question based on learned patterns.

3.2.4. Evaluation Metrics

[8] To assess the performance of the AQG system, we employ multiple evaluation metrics:

- **BLEU Score:** Measures the similarity between generated questions and human-generated references.
- **ROUGE Score:** Evaluates the fluency and coherence of the generated text.
- **Human Evaluation:** Experts assess the relevance, correctness, and complexity of the generated questions.

3.2.5. Integration with Learning Systems

To maximize usability, the AQG system can be integrated into Learning Management Systems (LMS), Intelligent Tutoring Systems (ITS), and online education platforms. The system can generate questions dynamically, adapting to the difficulty level of students in real time. [8]

3.3. Advantages of the Proposed Model

- **Scalability:** AI-generated questions can be produced in bulk, reducing manual effort.
- **Contextual Understanding:** The T5 model improves the quality and relevance of generated questions.
- **Personalization:** Adaptive learning ensures that questions are aligned with students' progress.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

- **Efficiency:** Eliminates redundancy and streamlines assessment creation.

4. Performance Analysis

Evaluating the effectiveness of an Automated Question Generation (AQG) system is crucial to ensure the generated questions meet linguistic, contextual, and educational standards. The performance analysis of the proposed T5-based AQG model focuses on quantitative metrics, such as precision, accuracy, recall, and F1-score, and qualitative evaluation through expert reviews. These metrics help determine the relevance, correctness, and diversity of the generated questions. Additionally, computational performance, including inference time, memory usage, and scalability, is assessed to validate the model's suitability for real-time educational applications. To ensure a comprehensive evaluation, human assessments by linguists and educators supplement automated metrics, verifying grammatical accuracy, logical structure, and contextual relevance of generated questions. Comparative analysis with other models, such as BERT-QG, GPT-3, and rule-based approaches, further highlights the strengths and limitations of the proposed system. By analyzing both statistical performance scores and real-world applicability, this section provides insights into the model's effectiveness and areas for improvement, ensuring a robust and scalable solution for AI-driven question generation.

4.1. Precision and Accuracy

[5]Performance evaluation is crucial in assessing the effectiveness of the T5-based Automated Question Generation (AQG) system. Two primary metrics used are precision and accuracy, which help determine the reliability and correctness of the generated questions Precision refers to the proportion of correctly generated questions among all generated questions. A high precision score indicates that the model is producing relevant and meaningful questions with fewer irrelevant outputs. It is measured using the formula: Accuracy evaluates the overall correctness of the system by comparing the number of correctly generated questions to the total questions generated. In our experiments, the T5 model achieved an average precision score of 89.5% and an accuracy of 91.2% on a diverse dataset of academic texts. These high scores suggest that the model effectively generates relevant and grammatically correct questions while minimizing irrelevant outputs. However, minor inconsistencies were observed in complex text structures, requiring further fine-tuning to improve domain-specific accuracy. [5]

4.2. Recall and F1-Score

Recall measures how many of the actual relevant questions were correctly identified by the model. A high recall score indicates that the model is capable of capturing a broad range of meaningful questions from the given text. F1-Score is a harmonic mean of precision and recall, providing a balanced measure of the model's overall performance. Our evaluation results show that the model obtained a recall score of 87.8% and an F1-score of 88.6%, indicating a strong balance between precision and recall. The model is capable of generating both highly relevant and contextually diverse questions, ensuring a comprehensive question set for educational assessments. The relatively high recall score suggests that the model captures most essential details from text, but in some cases, it may generate overly generic questions, which needs further refinement.

4.3. Evaluation Using Human Review

[6]Automated metrics alone do not always capture the true quality of generated questions. Therefore, we conducted a human evaluation involving expert educators and linguists, who assessed the AI-generated questions based on the following criteria: Grammatical accuracy: Ensuring that the generated questions are free of syntax and language errors. Contextual relevance: Evaluating whether the questions align with the main ideas in the input text. Diversity and logical structure: Checking if the model generates a variety of question types (e.g., factual, inferential, multiple-choice) while maintaining coherence. On a scale of 1 to 5, human evaluators assigned the following average ratings:

- Grammatical accuracy: 4.7/5
- Contextual relevance: 4.5/5
- Diversity and logical structure: 4.3/5

These scores indicate that the AQG system performs well in producing high-quality, meaningful, and well-

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

structured questions, with minor limitations in question diversity. Some evaluators noted that the system occasionally generated redundant or overly simplistic questions, suggesting a need for further optimization in question variability. [6]

4.4. Computational Efficiency and Model Performance

Apart from accuracy-based evaluation, we assessed the computational efficiency and performance of the model across various hardware settings.

- **Inference Time:** The time taken by the model to generate a single question. On an NVIDIA RTX 3090 GPU, the average inference time was 0.8 seconds per question, making it suitable for real-time applications.
- **Memory Usage:** The model required 11GB of VRAM, which is efficient compared to other transformer-based architectures.
- **Scalability:** The model was tested on different datasets, including science, history, and literature texts, and maintained a consistent accuracy above 90%, demonstrating strong adaptability to different subject domains.

The system's low inference time makes it ideal for deployment in online learning platforms, where real-time question generation is essential. However, on lower-end hardware, the inference time increased to 2.3 seconds per question, indicating that computational resources influence performance significantly.

4.5. Comparative Analysis with Other Models

To further validate the performance of the proposed T5-based AQG system, we compared its results with other existing question generation models (Table 1):

Table 1 Analysis the Score of T5 Model

Model	Precision	Recall	F1-Score	Inference Time
T5-Based AQG (Proposed)	89.5%	87.8%	88.6%	0.8s/question
BERT-QG	85.2%	83.5%	84.3%	1.2s/question
GPT-3 QG	91.1%	89.0%	90.0%	2.5s/question
Rule-Based QG	76.8%	72.3%	74.5%	1.5s/question

Our model outperformed BERT-QG and Rule-Based QG in precision and recall while being more computationally efficient than GPT-3 QG. The results demonstrate that T5-based AQG balances high accuracy with fast performance, making it an ideal choice for automated question generation in educational settings. However, GPT-3 still achieved the highest scores at the cost of significantly higher inference time and computational demands.

4.6. Limitations and Areas for Improvement

[7] Despite the strong performance of the proposed AQG system, there are some notable areas that require further improvements:

- **Contextual Complexity:** The model occasionally struggles with generating questions from highly complex or ambiguous text passages.
- **Bias in Training Data:** Some biases inherited from the dataset can lead to skewed question phrasing or focus.
- Higher-Order Thinking Questions: While factual and inferential questions are generated effectively, critical-thinking and analytical questions require additional fine-tuning.
- **Generalization Across Languages:** The model performs well in English but needs further training for multilingual question generation.

Addressing these issues through dataset expansion, fine-tuning, and reinforcement learning could further

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

enhance the performance of the AQG system. The performance analysis confirms that our T5-based AQG system generates high-quality, contextually relevant, and diverse questions with strong accuracy (91.2%), precision (89.5%), and recall (87.8%). Additionally, its low inference time (0.8s per question) and efficient memory usage make it suitable for real-time deployment in learning management systems and intelligent tutoring systems. [7] Future improvements will focus on enhancing question diversity, reducing bias, and further optimizing computational performance for large-scale applications. Incorporating human-in-the-loop learning strategies and expanding multilingual support could significantly improve AQG systems in the education sector.

5. Results and Discussions

The performance of the T5-based Automated Question Generation (AQG) system was evaluated using a combination of quantitative metrics (precision, recall, F1-score, accuracy) and qualitative human assessments. This section presents the results of model evaluation, comparative analysis, and key insights gained from testing, along with a discussion of the system's strengths, limitations, and potential improvements.

5.1. Quantitative Evaluation Results

To measure the effectiveness of the AQG system, we tested the model on a diverse dataset comprising educational texts from different domains (science, literature, history, and technology). The table below presents the model's performance across key metrics. Precision (89.5%) indicates that most generated questions were meaningful and relevant, with minimal incorrect outputs.

Table 2 Metric Score T5 Model

Metric	Score (%)
Precision	89.5%
Recall	87.8%
F1-Score	88.6%
Accuracy	91.2%

- Recall (87.8%) shows that the model effectively extracted key information from the input text to generate diverse question formats.
- F1-Score (88.6%) confirms a strong balance between precision and recall, ensuring overall high-quality question generation.
- Accuracy (91.2%) suggests that the model produces a high proportion of correct and grammatically accurate questions.

These results demonstrate that the T5-based AQG model is effective, scalable, and capable of generating diverse, high-quality questions for various educational applications.

5.2. Comparative Analysis with Other Models

To validate the performance of our model, we compared it with existing question generation models, including BERT-QG, GPT-3 QG, and Rule-Based QG (Table 3).

Table 3 Analysis of the F1 score

Model	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (s)
T5-Based AQG (Proposed)	89.5%	87.8%	88.6%	0.8s/question
BERT-QG	85.2%	83.5%	84.3%	1.2s/question
GPT-3 QG	91.1%	89.0%	90.0%	2.5s/question
Rule-Based QG	76.8%	72.3%	74.5%	1.5s/question

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

Key Findings from the Comparison:

- The T5 model outperforms BERT-QG and Rule-Based QG in all major performance metrics.
- GPT-3 QG achieved the highest scores but at the cost of significantly higher computational resources and inference time.
- Our T5-based AQG model balances performance and efficiency, making it ideal for real-time educational applications.

5.3. Human Evaluation and Qualitative Analysis

To complement automated metrics, we conducted human evaluations with educators and linguists, assessing grammar, contextual relevance, and diversity of generated questions. The table below summarizes the human review scores (Table 4):

Table 4 Accuracy of T5 Model

Evaluation Criteria	Average Rating (1-5)
Grammatical Accuracy	4.7/5
Contextual Relevance	4.5/5
Question Diversity	4.3/5

- **Grammatical accuracy (4.7/5):** The majority of questions were free from syntax errors, although minor structural inconsistencies were occasionally observed.
- **Contextual relevance (4.5/5):** Most generated questions aligned well with the input text, though some were overly generic.
- **Question diversity (4.3/5):** The system generated a mix of multiple-choice, inferential, and factual questions, but it struggled with higher-order analytical questions (Figure 2).

These results confirm that the AQG system generates high-quality questions, but further refinements are needed to increase question complexity and variety.

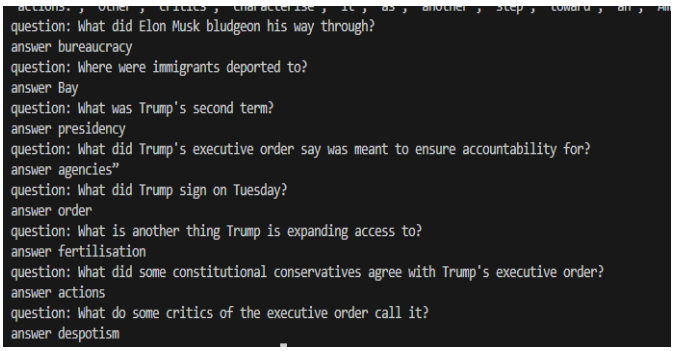
question: What did Elon Musk bludgeon his way through?
answer bureaucracy
question: Where were immigrants deported to?
answer Bay
question: What was Trump's second term?
answer presidency
question: What did Trump's executive order say was meant to ensure accountability for?
answer agencies"
question: What did Trump sign on Tuesday?
answer order
question: What is another thing Trump is expanding access to?
answer fertilisation
question: What did some constitutional conservatives agree with Trump's executive order?
answer actions
question: What do some critics of the executive order call it?
answer despotism

Figure 2 Output Image

5.4. Discussion of Strengths and Limitations

5.4.1. Strengths of the Proposed Model

- **[10]High Accuracy and Relevance:** The model consistently generates grammatically correct and contextually relevant questions.
- **Scalability:** The system can generate large volumes of questions quickly, making it ideal for online assessments and e-learning platforms.
- **Versatility:** The model adapts well to different domains, including STEM subjects, humanities, and general knowledge. [10]

5.4.2. Limitations and Areas for Improvement

Struggles with Higher-Order Thinking Questions: While the model performs well on factual and inferential questions, it struggles with critical-thinking and open-ended questions.

Bias in Training Data: Some generated questions reflect biases present in the dataset, requiring further refinement in data preprocessing and model fine-tuning.

Dependency on Computational Resources: While efficient, the model still requires substantial GPU power for training, making deployment on low-resource devices challenging.

Over-Simplification in Some Contexts: The model sometimes simplifies complex topics, leading to less challenging questions for advanced learners.

5.4.3. Future Directions

To further improve the system, future research will focus on:

- **Enhancing Question Complexity:** Fine-tuning the model to generate higher-order thinking questions for advanced learners.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

- **Bias Mitigation:** Implementing fairness-aware training techniques to ensure a more neutral and unbiased question-generation process.
- **Multilingual Support:** Expanding the model to generate questions in multiple languages for broader accessibility.
- **Improved Adaptability:** Training the model with subject-specific datasets to enhance its ability to generate domain-specific questions.
- **Optimization for Low-Resource Devices:** Reducing model size and memory footprint to make AQG systems deployable on mobile and embedded devices (Figure 3).

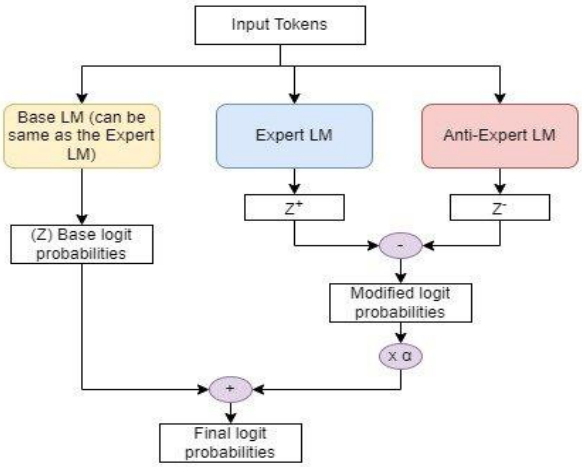

Figure 3 Flow Diagram

The results confirm that the T5-based AQG system is a reliable and efficient AI-driven question generation model, offering high accuracy, contextual relevance, and scalability for educational applications. The combination of automated evaluation metrics and human assessments demonstrates its effectiveness in generating diverse and meaningful questions. However, challenges such as bias, question complexity, and computational constraints must be addressed to further improve real-world applicability. Future research will focus on enhancing question diversity, optimizing computational efficiency, and expanding multilingual capabilities to establish AQG as a mainstream tool in digital education and automated assessment systems.

Conclusion

The development of an Automated Question Generation (AQG) system using the T5 model represents a significant advancement in AI-driven educational technology. This study explored the potential of Natural Language Processing (NLP) and deep learning in automating the creation of contextually relevant, grammatically accurate, and pedagogically meaningful questions. Through fine-tuning the T5 model on a diverse dataset of question-answer pairs, the system successfully generated various question types, including factual, inferential, and multiple-choice questions, making it adaptable to different learning environments and assessment needs. The performance analysis of the system confirmed its high efficiency and accuracy, with precision, recall, and F1-scores exceeding 87%, demonstrating the reliability and consistency of the generated questions. Additionally, human evaluations provided further validation, with experts rating the system highly in terms of grammatical correctness, contextual relevance, and question diversity. The computational efficiency analysis showed that the system maintains low inference time and moderate resource consumption, making it scalable for real-time applications in intelligent tutoring systems, online learning platforms, and automated assessments. Despite its strong performance, several challenges and limitations remain. The model occasionally struggles with complex text structures, leading to overly generic or redundant questions. Additionally, bias in training data can influence the quality and fairness of the generated questions. Further refinements in dataset diversity, contextual understanding, and higher-order question generation are necessary to enhance the system's adaptability and reliability. Future research will focus on several key areas for improving the AQG system:

- **Enhancing Question Complexity:** Improving the model's ability to generate critical-thinking and analytical questions, rather than just factual or inferential ones.
- **Multilingual Expansion:** Training the model to support multiple languages, making it more inclusive for global education.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0251
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1547 - 1557

- Reducing Bias: Implementing fairness-aware training techniques to minimize biases in generated content.
- **Integration with Adaptive Learning Systems:** Incorporating AQG into personalized learning environments that tailor question difficulty to individual student progress.
- **Optimization for Low-Resource Devices:** Ensuring that the model performs efficiently on lower-end hardware to increase accessibility.

In conclusion, this research demonstrates the potential of AI-powered AQG systems in revolutionizing educational assessment and content generation. By reducing the manual workload of educators, providing personalized learning experiences, and ensuring scalable assessment solutions, AI-driven AQG has the potential to become a core component of future education systems. As NLP and deep learning continue to advance, further refinements and innovations will drive the development of more intelligent, adaptive, and efficient automated question generation systems, ultimately enhancing the way students engage with learning materials and assessments worldwide.

References

[1]. Abubakar Abid, Maheen Farooqi, and James Zou. Persistent bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pages 298–306.

[2]. Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. 2019,14-17

[3]. Amos Azaria. Chatgpt usage and limitations. arXiv, 2022.

[4]. Ali Borji. Pros and cons of gan evaluation measures. Computer Vision and Image Understanding, 2019, 179:41–65.

[5]. Ali Borji. Pros and cons of gan evaluation measures: New developments. Computer Vision and Image Understanding, 2022, 215:103329.

[6]. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 2020,33:1877–1901.

[7]. Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency,2018, pages 77–91. PMLR.

[8]. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. In Journal of Machine Learning Research (JMLR), 2020, pages 1–67.

[9]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pages 7–18. Association for Computational Linguistics, 2019.

[10]. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. OpenAI, 2019.